NEW YORK UNIVERSITY
INSTITUTE OF MATHEMATICAL SCIENCES
LIBRARY
25 Waverly Place, New York 3, N. Y.

# AEC Computing and Applied Mathematics Center

## AEC RESEARCH AND DEVELOPMENT REPORT

TID-4500
14th Ed.

NYO-2881
PHYSICS

TWO METHODS USING POWER SERIES FOR SOLVING ANALYTIC INITIAL VALUE PROBLEMS

by

Glenn Lewis

January 1, 1960

NEW YORK UNIVERSITY
INSTITUTE OF MATHEMATICAL SCIENCES
LIBRARY
25 Waverly Place, New York 3, N. Y.

## Institute of Mathematical Sciences

NEW YORK UNIVERSITY
NEW YORK, NEW YORK

NYO-2881
c. 1

This report was prepared as an account of Government sponsored work. Neither the United States, nor the Commission, nor any person acting on behalf of the Commission:

A. Makes any warranty or representation, express or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or

B. Assumes any liabilities with respect to the use of, or for damages resulting from the use of any information, apparatus, method, or process disclosed in this report.

As used in the above, "person acting on behalf of the Commission" includes any employee or contractor of the Commission, or employee of such contractor, to the extent that such employee or contractor of the Commission, or employee of such cnntractor prepares, disseminates, or provides access to, any information pursuant to his employment or contract with the Commission, or his employment with such contractor.

AEC Computing and Applied Mathematics Center
Institute of Mathematical Sciences
New York University

TID-4500 NYO-2881
14th Ed. PHYSICS

# TWO METHODS USING POWER SERIES FOR SOLVING ANALYTIC INITIAL VALUE PROBLEMS

by

Glenn Lewis

January 1, 1960

Contract No. AT(30-1)-1480

NEW YORK UNIVERSITY
INSTITUTE OF MATHEMATICAL SCIENCES
LIBRARY
25 Waverly Place, New York 3, N. Y.

## ABSTRACT

Given any analytic non-characteristic initial value problem it is possible to obtain a truncated power series as an approximation to the solution. The complete power series, from which the truncation comes, may not have a domain of convergence large enough to cover the area of interest. Two methods for continuing these truncated power series, similar to analytic continuation for a complete power series, are given. Proofs are also given that these two methods of continuation converge, as the number of terms in the truncated power series goes to infinity, to the same values as the analytic continuation of the complete power series. These two methods are applied to the problem of finding the supersonic flow past a blunt axisymmetric body and the results are compared.

# TABLE OF CONTENTS

Page

Abstract . . . . . 2

Section

1. Introduction . . . . . 4
2. The First Method of Continuation . . . . . 12
3. The Second Method of Continuation . . . . . 23
4. Results of the Test Calculations . . . . . 34

Appendix A. Details of the Text Problem . . 49

Appendix B. The Arithmetic Used in the Test . . . . . 59

Table 1 . . . . . 62

Table 2 . . . . . 63

Table 3 . . . . . 65

Table 4 . . . . . 67

References . . . . . 68

# TWO METHODS USING POWER SERIES FOR SOLVING ANALYTIC INITIAL VALUE PROBLEMS

## 1. Introduction.

Over a period of years there has been a growing interest in problems in partial differential equations which are improperly posed (in the sense of Hadamard). Much of classical physics lead to problems in differential equations which are well posed such as the classical initial and boundary value problems. These well posed problems often admit of an elegant theoretical treatment as well as practical computational methods. This is not true of most of the problems which are improperly posed. These problems are often difficult to treat mathematically. Special cases have been treated by C. Pucci, M. Picone, F. John and others, but no general methods seem available. (See [1], [2], [3], and [4].) Many of these problems fall in the category of analytic non-characteristic initial value problems which are not hyperbolic. Even though the Cauchy-Kowalewski Theorem guarantees that there exists one and only one analytic solution for such problems (at least, in the small), it does not guarantee that other solutions do not exist, and even the analytic solutions need not depend continuously on the initial data. This paper is only

concerned, however, with finding the analytic solution to analytic non-characteristic initial value problems; non-analytic solutions, which probably cannot exist for such problems, are not considered.

One problem from fluid dynamics which, stated in an appropriate way, can be placed in this category is the supersonic flow past a blunt axisymmetric body. As is well known, a detached shock forms in front of the body. Ideally, it would be desirable to give the shape of the body and from this to determine the shape of the shock and the various flow quantities (fluid velocity, pressure, density, etc.) between it and the body. However, this is not an initial value problem and furthermore presents great mathematical difficulties. On the other hand if the shape of an axisymmetric shock is given as an analytic function then the flow quantities may be determined as analytic functions on the shock surface through the use of the Rankine-Hugoniot shock conditions. The equations of motion coupled with these conditions then presents an analytic initial value problem. The general characteristics of this flow are well known if a reasonable shape for the shock is assumed. The equations change type depending on the velocity of the flow. The flow is subsonic in a region in front of the tip of the body and the equations are elliptic;

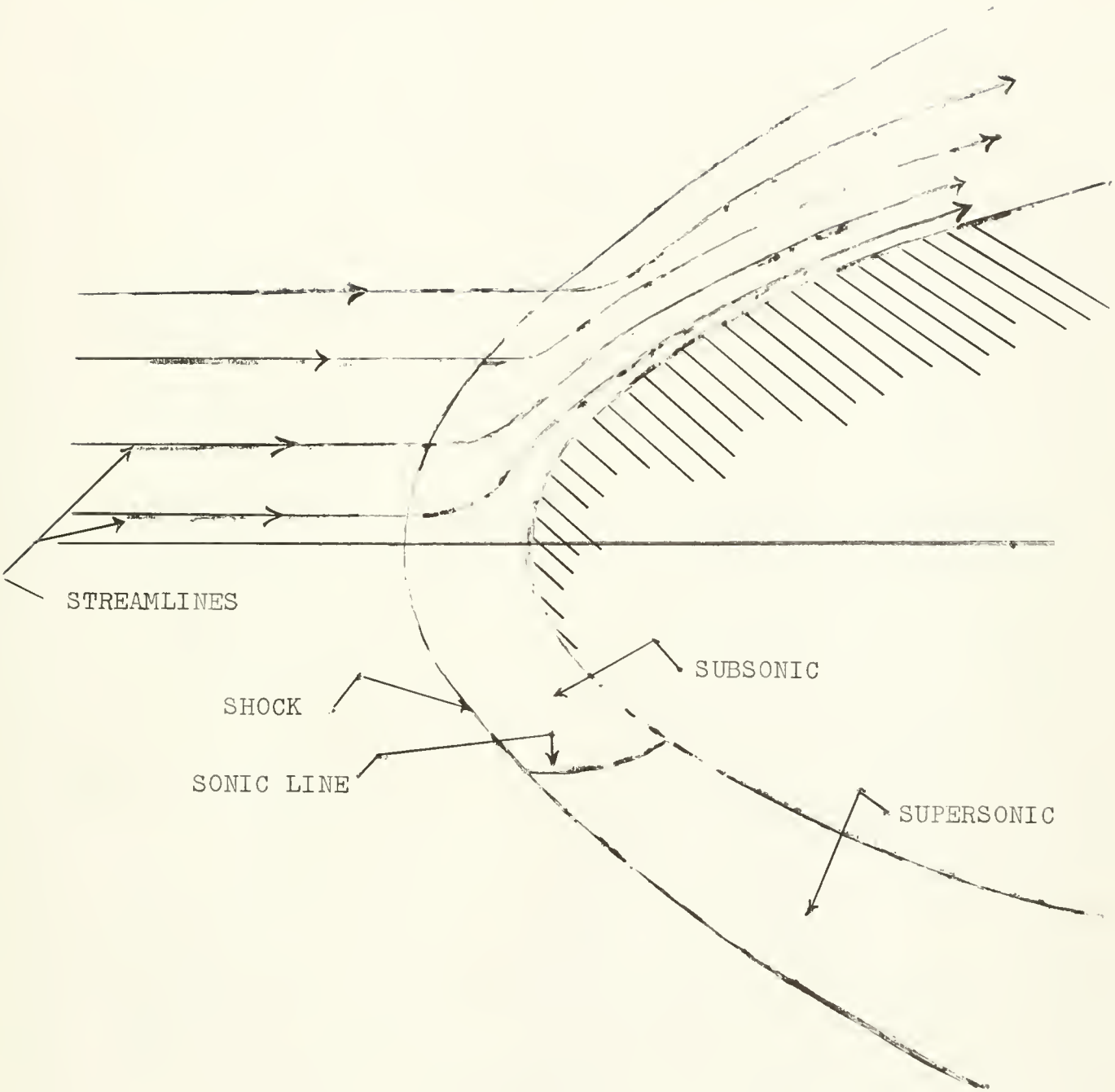


Figure 1

the flow becomes supersonic away from the tip and the equations become hyperbolic; see Figure 1, page 6. The problem is, therefore, of mixed type, and is probably improperly posed in the sense that continuous analytic perturbations of the shock shape will not produce continuous changes in the shape of the body. This may well be due to the concept of "improperly posed." Since this problem comes from physics, it is probably "properly posed" if a more restricted concept is used (e.g. only solutions which have an a priori bound in a given domain are compared with the original solution). All things considered, this problem makes an excellent test case for any computational method which claims to solve analytic initial value problems in general. The general characteristics are known, the constancy of entropy along streamlines and the constancy of the Bernoulli number afford exacting checks, and the problem is of practical as well as theoretical interest.

The first method which suggests itself in solving any analytic non-characteristic initial value problem is the use of power series. Following the method used in the proof of the Cauchy-Kowalewski Theorem, it is possible to compute all the low order coefficients of the solution expressed as a power series about any point on the initial surface. This truncated power

series should yield good approximations to the values of the solution at least near the point of expansion. This method was first applied to the detached shock problem by Lin and Rubinov (see [5]) and later by several others. This expansion was pursued furthest by R. D. Richtmyer (see [6]). He was the first to devise systematic recursion relations for the coefficients adequate for the use on a large-scale computer. (These relations are described in Appendix A.) His interest was primarily to investigate the use of power series for such problems as opposed to the interest of the fluid-dynamicists. He placed the problem on the Univac, carried nine figures of accuracy, and computed all coefficients of the flow quantities up to order 14. The results were excellent near the point of expansion, but there was some question as to whether the position of the body was accurately computed. To answer this question it was decided to recode the problem for the IBM 704 with increased accuracy. This report, then, is the direct outgrowth of that decision. In some sense, the present report is a continuation of R. D. Richtmyer's; but, for the convenience of the reader, it was decided to make the two reports independent.

The problem was recoded to carry 54 bits of accuracy,

and to handle coefficients up to order 19. It then became clear that the power series of the flow quantities diverged before the body was reached. This difficulty was predicted by M. D. Van Dyke (see [7]). Therefore, it became clear that a method for continuing truncated power series was needed similar to analytic continuation for the complete power series.

The obvious thing to do is to treat the truncated power series as if it were the complete power series and to analytically continue this truncated power series. Clearly, this is doomed to fail. The truncated power series is but a polynomial, say in x of order k, and analytic continuation about a point $x_o$ of this polynomial simply expresses this polynomial in powers of $(x-x_o)$ and does not change its values anywhere. As k goes to infinity (i.e. more and more terms of the original power series are taken), the resulting polynomials must diverge for any point which is neither in nor on the boundary of the domain of convergence of the original power series. This is merely a restatement of the fact that a power series must diverge for any point which is strictly outside of its domain of convergence. However, if a certain proportion of the high order terms of the polynomials expressed in powers of $(x-x_o)$ are, discarded, the approximations will converge to the

desired value as k goes to infinity. This is called the first method of continuation, is described in detail in Section II, and a proof of convergence as k goes to infinity is given.

The second method of continuation described in detail in Section III uses a different approach. As in the first method, a truncated power series is computed on the initial surface; but, now, this truncated power series is used to approximate initial data on a new initial surface. From this new initial data a new truncated power series is computed using the differential equations again. If the new initial surface satisfies certain restrictions and if the order of the new truncated power series is a certain fraction of the order of the old truncated power series, then it is shown in Section III that this new truncated power series converges as the number of terms goes to infinity for arguments outside of the domain of the original power series. Surprisingly enough, the second method was the first discovered. Various methods were discussed, and finally the second method was tried on the machine. The results were extremely good. In attempting to justify these results the much simpler first method was hit upon.

Section IV describes and compares the results

when the two methods were applied to the detached shock problem. Appendix A describes the detached shock problem in detail and gives the details of the actual test computations, and Appendix B describes the unnormalized double-precision floating-point arithmetic used.

The author would like to express his gratitude to Dr. R. D. Richtmyer, Prof. E. Isaacson, and Prof. F. John whose continued interest and helpful suggestions made this paper possible.

## 2. The First Method of Continuation.

Description and proof of convergence. As was already stated, it is possible to compute the low order coefficients of the analytic solution of an analytic initial value problem. (The actual computation of these coefficients is discussed in detail in Appendix A.) In this section the solution will be thought of as an analytic function of which the low order coefficients of its power series are known and no further reference to the analytic initial value problem will be made.

Let $F(x_i)$ be an analytic function of the variables $x_1,\ldots,x_q$ which is expressible as a power series about the origin. Let $D_o$ be the domain of convergence of this power series (that is, the largest open set which contains only points for which the power series is absolutely convergent). Let $x_i^o$ be any fixed point in $D_o$. Under the assumptions $F(x_i)$ may be expressed as a power series about the point $x_i^o$. Let $D_1$ be the domain of convergence of this new power series, and let $x_i^1$ be any fixed point in $D_1$. The problem, then, is to find an approximation to $F(x_i^1)$ when only the truncated power series about the origin of $F(x_i)$ is known.

Introduce the new coordinate system $\xi_i = x_i - x_i^o$. Let $f(\xi_i) = F(\xi_i + x_i^o)$ and let $\xi_i^o = x_i^1 - x_i^o$, then $F(x_i^1) = f(\xi_i^o)$.

Assuming only the terms of order less than or equal to k of $F(x_i)$ expanded about the origin are known, let $F_k(x_i)$ be this truncated power series, and let $f_k(\xi_i) = F_k(\xi_i + x_i^o)$. Finally let $f_{s,k}(\xi_i)$ be the terms of $f_k(\xi_i)$ of order less than or equal to [sk] where s is some constant between zero and one ([sk] is the greatest integer in sk). That is, the truncated power series $F_k(x_i)$ is written in terms of the new coordinate system $\xi_i$, and then only terms of $\xi_i$ of order less than or equal to [sk] are retained to obtain the final approximation.

Theorem: Under the above assumptions, there exists an s, $0 < s < 1$, such that $f_{s,k}(\xi_i^o) \to F(x_i^1)$ as $k \to \infty$.

Before the proof it is convenient to consider the properties of the function g(z),

$$g(z) = AB^z \frac{1}{z^z(1-z)^{1-z}} ,$$

for $0 \leq s \leq 1$ where A and B are positive constants. The minima of this function occur at the extremes of the range: $g(z) = A$ when $z = 0$, $g(z) = AB$ when $z = 1$. The maximum, $g(z) = A(B+1)$, occurs when $z = B/(1+B)$. Thus the equation, $g(z) = 1$, has at most two roots, see Figure 2, page 15.

Choose $R_i$ such that $R_i$ is in $D_o$ and $|x_i^o| < R_i$.

This is always possible, see [9]. Let $A = \max_i |x_i^o|/R_i$ and $B = \max_i |\xi_i^o|/|x_i^o|$. Choose an s, $0 < s < 1$, such that

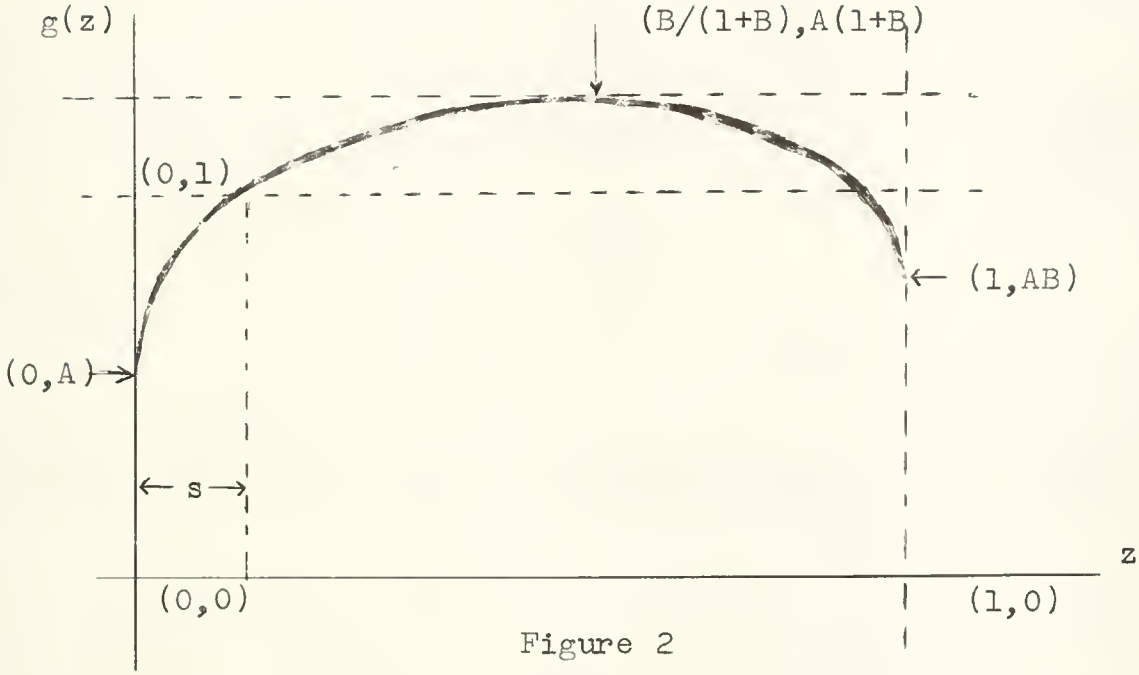


Figure 2

s is smaller than the roots of $g(z) = 1$. Under this assumption it will be established in the proof that $f_{s,k}(\xi_i^o) \to F(x_i^1)$ as $k \to \infty$. Since $A < 1$, it is always possible to choose an s which satisfies the above property.

Proof: Since $F(x_i)$ is assumed to be analytic at the origin, it has a power series expansion:

$$F(x_i) = \sum_{n=0}^{\infty} \sum_{n_1+\ldots+n_q=n} a_{n_1,\ldots,n_q} x_1^{n_1} \ldots x_q^{n_q} .$$

(All subscripts are assumed to be non-negative.) $F(x_i^1)$ may therefore be written in the following way:

$$F(x_i^1) = \sum_{m=0}^{\infty} \sum_{m_1+\ldots+m_q=m} b_{m_1,\ldots,m_q} \xi_1^{o\,m_1} \ldots \xi_q^{o\,m_q}$$

where $b_{m_1,\ldots,m_q}$ equals

$$\sum_{n=0}^{\infty} \sum_{n_1+\ldots+n_q=n} \binom{n_1}{m_1} \ldots \binom{n_q}{m_q} a_{n_1,\ldots,n_q} x_1^{o\,n_1-m_1} \ldots x_q^{o\,n_q-m_q}.$$

$$\left( \binom{n}{m} = 0 \text{ if } n < m \text{ or } m < 0 \right).$$

Let $b^k_{m_1,\ldots,m_q}$ equal

$$\sum_{n=0}^{k} \sum_{n_1+\ldots+n_q=n} \binom{n_1}{m_1} \ldots \binom{n_q}{m_q} a_{n_1,\ldots,n_q} x_1^{o\,n_1-m_1} \ldots x_q^{o\,n_q-m_q}$$

then

$$f_{s,k}(\xi_i^o) = \sum_{m=0}^{[sk]} \sum_{m_1+\ldots+m_q=m} b^k_{m_1,\ldots,m_q} \xi_1^{o\,m_1} \ldots \xi_q^{o\,m_q}.$$

The error in the approximation, i.e. $F(x_i^1)-f_{s,k}(\xi_i^o)$, can be separated into two kinds. Part of the error comes from the fact that the approximation is a polynomial and that the coefficients of order greater than [sk] are zero. This is called the error of the first kind and is denoted by $E_1$. The rest of the error comes from the fact that the coefficients of this polynomial approximation differ from the corresponding coefficients of the exact solution. This is called the error of the second kind and is denoted by $E_2$. The error obtained

when applying the second method of continuation, described in Section III, can also be divided into errors of the first and second kind. From this it follows that $F(x_i^1) = f_{s,k}(\xi_i^o)+E_1+E_2$ where

$$E_1 = \sum_{m=[sk]+1}^{\infty} \sum_{m_1+\ldots+m_q=m} b_{m_1,\ldots,m_q} \xi_1^{o\ m_1} \ldots \xi_q^{o\ m_q}$$

and

$$E_2 = \sum_{m=0}^{[sk]} \sum_{m_1+\ldots+m_q=m} (b_{m_1,\ldots,m_q} - b^k_{m_1,\ldots,m_q}) \xi_1^{o\ m_1} \ldots \xi_q^{o\ m_q}.$$

Since $E_1$ is just the tail end of the expansion for $f(\xi_i^o)$, by assumption, it will converge to zero. Therefore the problem is to show that $E_2 \to 0$ as $k \to \infty$. From the assumptions on the domain of convergence of $F(x_i)$, it follows that we may use Cauchy's Integral Formula over and over to obtain the inequality on the coefficients:

$$|a_{n_1,\ldots,n_q}| \leq \frac{M}{R_1^{n_1} \ldots R_q^{n_q}}$$

where M is the maximum value of $F(x_i)$ for $|x_i| \leq R_i$, see [10]. Using this, it follows that $|E_2|$ is less than or equal to

$$\sum_{m=0}^{[sk]} \sum_{m_1+\ldots+m_q=m} \sum_{n=k+1}^{\infty} \sum_{n_1+\ldots+n_q=n} M\binom{n_1}{m_1}\ldots\binom{n_q}{m_q} A^n B^m .$$

It is well known that

$$\sum_{m_1+\ldots+m_q=m} \binom{n_1}{m_1}\ldots\binom{n_q}{m_q} = \binom{n_1+\ldots+n_q}{m} . \tag{1}$$

(This may be found in most texts on elementary probability.) The following is also true:

$$\sum_{n_1+\ldots+n_q=n} 1 = \binom{n+q-1}{n} . \tag{2}$$

(The binomial coefficient may easily be shown by induction on n to satisfy the same recursion relation that the sum does.) Putting (1) and (2) together we obtain:

$$|E_2| \leq \sum_{m=0}^{[sk]} \sum_{n=k+1}^{\infty} M\binom{n}{m}\binom{n+q-1}{n} A^n B^m . \tag{3}$$

Case one: It is possible that $g(z) < 1$ for all $0 \leq z \leq 1$. In this case the approximation will converge for any s, $0 < s \leq 1$. Since $g(z) < 1$, $A(B+1) < 1$. But this implies that $|x_i^o|+|x_i^1-x_i^o| < R_i$ for all i or $|x_i^1| < R_i$; that is, $x_i^1$ is in $D_o$, and the case is essentially trivial. If we let the m index run from 0 to n we obtain:

$$\begin{aligned} |E_2| &\leq \sum_{n=k+1}^{\infty} M\binom{n+q-1}{n} A^n \sum_{m=0}^{n} \binom{n}{m} B^m \\ &\leq \sum_{n=k+1}^{\infty} M\binom{n+q-1}{n}(A(1+B))^n . \end{aligned}$$

The expression on the right is clearly the tail end of a power series which converges if $A(1+B) < 1$. Hence, in this case, $E_2 \to 0$ as $k \to \infty$.

Case two: $g(z) = 1$ for some $z$, $0 < z < 1$. Since the maximum of $g(z)$ is achieved at $z = B/(1+B)$,

$$s < B/(1+B) . \tag{4}$$

The assumption about $s$ implies $g(s) < 1$. From the inequality $g(s) < 1$ and (4) it follows that

$$s < 1-A . \tag{5}$$

Rearranging terms, we may write (3) in the following way:

$$|E_2| \leq M \sum_{n=k+1}^{\infty} A^n \binom{n+q-1}{n} \sum_{m=0}^{[sk]} \binom{n}{m} B^m .$$

Consider the expression $\sum_{m=0}^{[sk]} \binom{n}{m} B^m$. $\binom{n}{[sk]} B^{[sk]}$ will be the largest term in this sum since

$$\binom{n}{[sk]-1} B^{[sk]-1} \leq \binom{n}{[sk]} B^{[sk]}$$

or

$$\frac{[sk]}{n-[sk]+1} \leq B .$$

This follows immediately since, by (4), $B > 1/(1-s)$ and $n \geq k+1$ implies $1/(1-s) > [sk]/(n-[sk]+1)$. Therefore it follows that

$$\sum_{m=0}^{[sk]} \binom{n}{m} B^m \leq (sk+1)\binom{n}{[sk]} B^{[sk]}$$

and

$$|E_2| \leq M(sk+1)B^{[sk]} \sum_{n=k+1}^{\infty} \binom{n}{[sk]}\binom{n+q-1}{n} A^n .$$

Changing the index of summation this may be rewritten:

$$|E_2| \leq M(sk+1)B^{sk}\binom{k+1}{sk}\binom{k+q}{k+1} A^{k+1} \sum_{n=0}^{\infty} WA^n$$

where

$$W = \frac{\binom{n+k+1}{sk}\binom{n+k+q}{n+k+1}}{\binom{k+1}{sk}\binom{k+q}{k+1}} .$$

[sk] has been replaced by sk; this simplifies the notation and it does not change the analysis.

$$W = \frac{(k+1 \quad +1)\ldots(k+1 \quad +n)(k+q+1)\ldots(k+q+n)}{(k+1-sk+1)\ldots(k+1-sk+n)(k+1+1)\ldots(k+1+n)}$$

$$\leq \left(\frac{1}{1-s}\right)^n \left(\frac{k+q}{k+1}\right)^n .$$

This follows, since $\frac{k+n}{k-sk+n} < \frac{1}{1-s}$ and $\frac{k+q+n}{k+1+n} < \frac{k+q}{k+1}$ for all $n > 0$. Hence

$$|E_2| \leq M(sk+1)B^{sk}\binom{k+1}{sk}\binom{k+q}{k+1} A^{k+1} \sum_{n=0}^{\infty} \left(\frac{A}{1-s} \cdot \frac{k+q}{k+1}\right)^n .$$

But the infinite sum converges since (5) guarantees that $\frac{A}{1-s} \cdot \frac{k+q}{k+1} < 1$ if k is large enough. Therefore

$$|E_2| \leq M(sk+1)B^{sk}\binom{k+1}{sk}\binom{k+q}{k+1} A^{k+1} \frac{1}{1-\frac{A(k+q)}{(1-s)(k+1)}} .$$

Using Sterling's Approximation to the binomial coefficient $\binom{k+1}{sk}$, we finally obtain:

$$|E_2| \leq \frac{M(sk+1)}{\sqrt{2\pi sk(1-s)}} \binom{k+q}{k+1} \frac{(g(s))^k}{1-\frac{A(k+q)}{(1-s)(k+1)}} .$$

But $\binom{k+q}{k+1}$ is just a polynomial in k, therefore the whole expression will be dominated by $g(s))^k$. Since $g(s) < 1$, $E_2 \to 0$ as $k \to \infty$.

It should be noted here that $E_2$ converges to zero under the assumptions that $x_o$ is in $D_o$ and that s satisfies the condition that s is less than the roots of the equation $g(z) = 1$. It is not necessary for the convergence of $E_2$ that $x_1$ is in $D_1$. This fact will be used later in showing that the second method of continuation converges.

<u>Concerning Repeated Continuations</u>. Values of an analytic function obtained through repeated analytic continuation may be approximated by truncated power series. For simplicity the method will be stated for analytic functions of only one variable. Since the proof follows closely that which was already given, just the statement follows.

Let f(x) be an analytic function given as a series in powers of x. Let $x_1, \ldots, x_q$ be a sequence of points about which the power series may be analytically continued, i.e. the power series may first be rewritten in powers of $x-x_1$, this new power series may be rewritten in powers of $x-x_2$, etc.. Using the sequence of continuations indicated, the problem is then to find an approximation to $f(x_q)$.

Call the origin $x_0$. Then, let $R_1$ be the radius of convergence of the power series about $x_0$, let $R_2$ be the radius of convergence of the power series about $x_1$, etc.. From the assumptions about how the function may be analytically continued, it follows that $|x_i - x_{i-1}| < R_i$, for $i = 1$ to $q$. It is assumed that only the terms of order less than or equal to $k$ of the power series about the origin are known. Let $s_i$, $i = 1$ to $q-1$, be a sequence of real numbers, $0 < s_i < 1$. The k'th order polynomial, which is an approximation to the power series about the origin, is rewritten in terms of powers of $x-x_1$, and terms of order greater than $[s_1 k]$ are discarded. This new polynomial is rewritten in terms of powers of $x-x_2$, and terms of order greater than $[s_1 s_2 k]$ are discarded, etc.. This process is repeated until a polynomial of order $[s_1 \ldots s_{q-1} k]$ in powers of $x-x_{q-1}$ is obtained. Then $x_q$ is substituted for $x$. If the $s_i$'s are properly chosen, the above approximation will converge to $f(x_q)$ as $k \to \infty$.

Assuming the $s_i$'s are given. define the $A_i$'s by induction:

$$A_q = 1$$

$$A_i = \frac{1}{s_i^{s_i}(1-s_i)^{1-s_i}} \left|\frac{x_{i+1}-x_i}{x_i-x_{i-1}}\right|^{s_i} A_{i+1}^{s_i} .$$

If

$$\frac{s_i}{1-s_i} < \left|\frac{x_{i+1}-x_i}{x_i-x_{i-1}}\right|$$

$$s_i < 1 - \frac{|x_i-x_{i-1}|}{R_i}$$

and if $A_i \frac{|x_i-x_{i-1}|}{R_i} < 1$ for $i = 1,\ldots,q-1$,

then the approximation converges as $k \to \infty$.

Furthermore, given any set of $x_i$'s, and $R_i$'s, there exist $s_i$'s which satisfy the above conditions.

## 3. The second Method of Continuation.

Description of the Method. This method makes more use of the differential equations than the first. In order to simplify notation the method will be described and the proof of convergence will be given for an analytic initial value problem in only two variables and one unknown. With only trivial modifications the method and proof are applicable to the most general analytic non-characteristic initial value problem.

Let
$$\frac{\partial u}{\partial x} = G(u)\,\frac{\partial u}{\partial y} \tag{6a}$$
and
$$u(0,y) = f(y),\quad f(0) = 0 \tag{6b}$$
where $G(u)$ is analytic at $u = 0$, $G(u) = \sum_{n=0}^{\infty} a_n u^n$,
and $f(y)$ is analytic at $y = 0$, $f(y) = \sum_{n=0}^{\infty} b_n y^n$,
and only the coefficients of order less than or equal to $k$ of this expansion of $f(y)$ are known. Let $u_k(x,y)$ be the polynomial which is obtained from $u(x,y)$ when $u(x,y)$ is expanded in a series in powers of $x$ and $y$ and terms of order greater than $k$ are discarded.

From the differential equation (6a) it is possible to write the partial derivatives of $u$ of any order in terms of partial derivatives of $u$ with respect to $y$ of that order or less by differentiating over and over again. Therefore, using (6b), it is possible to compute $u_k(x,y)$ when only the coefficients of $f(y)$ of order less than or equal to $k$ are known.

Let $(x_o, y_o)$ be a point in the domain of absolute convergence of the power series of $u(x,y)$ expanded about the origin (the choice of this point will be further restricted to insure convergence of the method) and let $(x_1, y_1)$ be a point in the domain of absolute convergence of the power series of $u(x,y)$ expanded about the point $(x_o, y_o)$. The problem then is to find an approximation to $u(x_1, y_1)$.

Introduce the new coordinate system $\xi = x - x_o$ and $\zeta = y - y_o$, and let $\xi_o = x_1 - x_o$ and $\zeta_o = y_1 - y_o$.

Let $$\frac{\partial w}{\partial \xi} = G(w) \frac{\partial w}{\partial \zeta} \tag{7a}$$

and $$w(0, \zeta) = u_k(x_o, \zeta + y_o) \tag{7b}$$

and finally, let $w_{[sk]}(\xi, \zeta)$ be the polynomial which is obtained from $w(\xi, \zeta)$ when $w(\xi, \zeta)$ is expanded in powers of $\xi$ and $\zeta$ and all powers of order greater than $[sk]$ are discarded where $s$ is some constant between zero and one. ($[sk]$ is the greatest integer in $sk$.) If equation (7b) were replaced by $w(0, \zeta) = u(x_o, \zeta + y_o)$ then $w(\xi_o, \zeta_o)$ would equal $u(x_1, y_1)$. As it stands, we shall show that $w_{[sk]}(\xi_o, \zeta_o) \to u(x_1, y_1)$ as $k \to \infty$ if $s$ and $(x_o, y_o)$ satisfy certain restrictions.

These restrictions on $s$ and $(x_o, y_o)$ can be stated in terms of the following dominating problem:

$$\frac{\partial v}{\partial x} = \overline{G}(v)\,\frac{\partial v}{\partial y} \tag{8a}$$

and

$$v(0,y) = \overline{f}(y) \tag{8b}$$

where $\overline{G}(v) = \sum_{n=0}^{\infty} |a_n|\, v^n$ and $\overline{f}(y) = \sum_{n=0}^{\infty} |b_n| y^n$ .

The term "dominating problem" requires some explanation. Let $\phi(\xi,\zeta) = \sum c_{m,n}\xi^n\zeta^n$ and $\psi(\xi,\zeta) = \sum d_{m,n}\xi^m\zeta^n$. $\phi(\xi,\zeta)$ is said to dominate $\psi(\xi,\zeta)$, $\phi(\xi,\zeta) \supseteq \psi(\xi,\zeta)$, if $c_{m,n} \geq |d_{m,n}|$. A problem (a differential equation plus an initial condition in the same form as (6a) and (6b)) is said to dominate another problem if the functions occuring in the former dominate the corresponding functions occuring in the latter. This concept is useful for it is easy to show that the solution of the dominating problem dominates the solution of the other problem. Since $\overline{G}(u) \supseteq G(u)$ and $\overline{f}(y) \supseteq f(y)$, problem (8) dominates problem (6), hence $v(x,y) \supseteq u(x,y)$. This concept of domination is the most important tool in the proof of convergence.

Theorem: If $(x_o,y_o)$ is in the domain of convergence of the power series of $v(x,y)$ expanded about the origin and if $(x_1,y_1)$ is in the domain of convergence of the power series of $u(x,y)$ expanded about the point $(x_o,y_o)$ then there exists an s such that $w_{[sk]}(\xi_o,\zeta_o) \to u(x_1,y_1)$ as $k \to \infty$.

Since $(x_o,y_o)$ is in the domain of convergence

of the power series of $v(x,y)$ expanded about the origin, there exist two positive numbers, $X$ and $Y$, such that $(X,Y)$ is in this domain and $|x_o| < X$ and $|y_o| < Y$.

Let $A = \max\left(\frac{|x_o|}{X}, \frac{|y_o|}{Y}\right)$, and $B = \max\left(\frac{|\xi_o|}{|x_o|}, \frac{|\zeta_o|}{|y_o|}\right)$,

and
$$g(z) = AB^z \frac{1}{z^z(1-z)^{1-z}} .$$

If $s$ is chosen such that it is positive and less than the roots of $g(z) = 1$, then the approximation will converge.

The domain of absolute convergence of the power series of $v(x,y)$ expanded about the origin may well be properly included in the corresponding domain of $u(x,y)$. The restricting of $(x_o, y_o)$ to this possibly smaller domain is unnecessary in the proof of the convergence of the first method of continuation and seems somewhat arbitrary in the present case. (An example has been found for which this restriction is unnecessary.) However, since the Cauchy-Kowalewski Theorem guarantees that this possibly smaller domain is not empty, the second method of continuation will converge for a larger domain than the domain of absolute convergence of the power series of $u(x,y)$ expanded about the origin unless the boundary of this domain is the natural boundary of $u(x,y)$. That is, even though the second method of

continuation can only be shown to converge in a domain which may be smaller than the domain of convergence of the first method, this domain always contains points for which the functions could not be approximated unless some method of continuation were used.

Proof of Convergence. It is convenient to introduce five new operators, S, I, $\overline{I}$, T, and R, and then to describe the approximation in terms of these operators.

The operation of substituting $x_o$ for the variable x in a function $\phi(x)$ is indicated by $S_x^{x_o}\phi(x)$; that is, $S_x^{x_o}\phi(x) = \phi(x_o)$.

The operator I (I for initial value problem) may only be applied to a function of $\zeta$. By $I\phi(\zeta)$ is meant the unique analytic function of the variables $\xi$ and $\zeta$ which solves the partial differential equation (7a) and is equal to $\phi(\zeta)$ when $\xi = 0$. That is, $\psi(\xi,\zeta) = I\phi(\zeta)$ if and only if $\frac{\partial\psi}{\partial\xi} = G(\psi)\frac{\partial\psi}{\partial\zeta}$ and $\psi(0,\zeta) = \phi(\zeta)$. The operator $\overline{I}$ is the same as I except the differential equation $\frac{\partial\psi}{\partial\xi} = \overline{G}(\psi)\frac{\partial\psi}{\partial\zeta}$ is used instead of (7a).

The operator $T_j^{x,y}$ (T for truncate) can only be applied to an analytic function, $\phi(x,y)$, which is analytic at the point $x = 0$, $y = 0$. By $T_j^{x,y}\phi(x,y)$ is meant the polynomial in x and y which is obtained when $\phi(x,y)$ is expanded in powers of x and y and terms of order greater than j are discarded. That is, if

$$\phi(x,y) = \sum_{q=0}^{\infty} \sum_{m+n=q} a_{m,n}\, x^m\, y^n$$

then

$$T_j^{x,y}\phi(x,y) = \sum_{q=0}^{j} \sum_{m+n=q} a_{m,n}\, x^m\, y^n \ .$$

$T_j^{\xi,\zeta}$, $T_j^{x,y,x_o,y_o}$, etc. are defined in a similar manner.

The operator R (R for remainder) may be applied to the same functions that T can. $R_j^{x,y}\phi(x,y)$ is defined to be equal to $\phi(x,y) - T_j^{x,y}\phi(x,y)$, etc..

Let $u(x,y)$ be the solution of (6) and let $U(\xi,\zeta) = u(\xi+x_o,\zeta+y_o)$. Let $v(x,y)$ be the solution of (8) and let $V(\xi,\zeta) = v(\xi+|x_o|,\zeta+|y_o|)$.

Let $\quad v_1(\xi,\zeta) = S_x^{\xi+|x_o|}(S_y^{\zeta+|y_o|}(T_k^{x,y}v(x,y)))$,

$$v_2(\xi,\zeta) = \bar{I}(S_x^{|x_o|}(S_y^{\zeta+|y_o|}(T_k^{x,y}v(x,y)))),$$

$$u_2(\xi,\zeta) = I(S_x^{x_o}(S_y^{\zeta+y_o}(T_k^{x,y}u(x,y)))) \ .$$

$S_\xi^{\xi_o}S_\zeta^{\zeta_o}T_{[sk]}^{\xi,\zeta}u_2(\xi,\zeta)$ is the approximation already described. The problem is then to show that

$$S_\xi^{\xi_o}S_\zeta^{\zeta_o}T_{[sk]}^{\xi,\zeta}u_2(\xi,\zeta) \to U(\xi_o,\zeta_o) \quad \text{or} \quad u(x_1,y_1)$$

as $k \to \infty$. $u(x_1,y_1)$ can be written in the following way:

$$u(x_1,y_1) = S_\xi^{\xi_o} S_\zeta^{\zeta_o} T_{[sk]}^{\xi,\zeta}(U(\xi,\zeta) - u_2(\xi,\zeta))$$

$$+ S_\xi^{\xi_o} S_\zeta^{\zeta_o} R_{[sk]}^{\xi,\zeta} U(\xi,\zeta)$$

$$+ S_\xi^{\xi_o} S_\zeta^{\zeta_o} T_{[sk]}^{\xi,\zeta} u_2(\xi,\zeta) \ .$$

Since $(x_1,y_1)$ was assumed to be in the domain of convergence of the power series of $u(x,y)$ expanded about the point $(x_o,y_o)$, this implies that $(\xi_o,\zeta_o)$ is in the domain of convergence of the power series of $U(\xi,\zeta)$ expanded in powers of $\xi$ and $\zeta$. $S_\xi^{\xi_o} S_\zeta^{\zeta_o} R_{[sk]}^{\xi,\zeta} U(\xi,\zeta)$, the error of the first kind as described in Section II, is just the tail end of the power series for $U(\xi_o,\zeta_o)$, and therefore by assumption converges to zero. Therefore the problem is reduced to showing that the error of the second kind converges to zero, i.e.

$$S_\xi^{\xi_o} S_\zeta^{\zeta_o} T_{[sk]}^{\xi,\zeta}(U(\xi,\zeta) - u_2(\xi,\zeta)) \to 0 \text{ as } k \to \infty .$$

This is easily established by the use of two simple lemmas.

Lemma 1: $V(\xi,\zeta) - v_2(\xi,\zeta) \geq U(\xi,\zeta) - u_2(\xi,\zeta)$.

Let $M(\zeta,\delta) = S_x^{|x_o|} S_y^{\zeta+|y_o|}(T_k^{x,y} v(x,y) + \delta R_k^{x,y} v(x,y))$

and $N(\zeta,\delta) = S_x^{x_o} S_y^{\zeta+y_o}(T_k^{x,y} u(x,y) + \delta R_k^{x,y} u(x,y))$.

Since $v(x,y) \geq u(x,y)$, it follows that

$$M(\zeta,\delta) \geq N(\zeta,\delta) \ . \tag{9}$$

The differential equations associated with I and $\bar{I}$ can be thought of as containing $\delta$ vacuously; therefore, $\bar{I}(M(\zeta,\delta)) \supseteq I(N(\zeta,\delta))$ with respect to the variables $\xi$, $\zeta$, and $\delta$. Since $S_\delta^0\, \bar{I}(M(\zeta,\delta)) = v_2(\xi,\zeta)$ and $S_\delta^0\, I(N(\zeta,\delta)) = u_2(\xi,\zeta)$, there exists two analytic functions $M_1(\xi,\zeta,\delta)$ and $N_1(\xi,\zeta,\delta)$ such that

$$\bar{I}\ (M(\zeta,\delta) = v_2(\xi,\zeta) + \delta M_1(\xi,\zeta,\delta) \tag{10}$$

and

$$I(N(\zeta,\delta)) = u_2(\xi,\zeta) + \delta N_1(\xi,\zeta,\delta) \tag{11}$$

From (9) it follows that $M_1(\xi,\zeta,\delta) \supseteq N_1(\xi,\zeta,\delta)$. This would still be true of 1 (one) were substituted for $\delta$. Substituting 1 for $\delta$ in equations (10) and (11) we obtain:

$$V(\xi,\zeta) = v_2(\xi,\zeta) + M_1(\xi,\zeta,1)$$

$$U(\xi,\zeta) = u_2(\xi,\zeta) + N_1(\xi,\zeta,1)$$

Therefore $V(\xi,\zeta) - v_2(\xi,\zeta) \supseteq U(\xi,\zeta) - u_2(\xi,\zeta)$.

Lemma 2: $V(\xi,\zeta) \supseteq v_2(\xi,\zeta) \supseteq v_1(\xi,\zeta) \supseteq 0$ .

Since $v(|x_o|,\zeta+|y_o|) \supseteq S_x^{|x_o|} S_y^{\zeta+|y_o|} T_k^{x,y} v(x,y)$, it follows that

$$\bar{I}(v(|x_o|,\ \zeta+|y_o|)) \supseteq \bar{I}(S_x^{|x_o|} S_y^{\zeta+|y_o|} T_k^{x,y} v(x,y)).$$

But $\quad V(\xi,\zeta) = \bar{I}(v(|x_o|,\zeta+|y_o|))$

and $\quad v_2(\xi,\zeta) = \bar{I}(S_x^{|x_o|} S_y^{\zeta+|y_o|} T_k^{x,y} v(x,y))$,

hence $$V(\xi,\zeta) \geq v_2(\xi,\zeta) \ .$$

$$T_k^{\xi,\zeta,|x_o|,|y_o|} v_2(\xi,\zeta) = T_k^{\xi,\zeta,|x_o|,|y_o|} \overline{I}(S_x^{|x_o|} S^{\zeta+|y_o|} T_k^{x,y} v(x,y))$$

$$= T_k^{\xi,\zeta,|x_o|,|y_o|} \overline{I}(T_k^{\zeta,|x_o|,|y_o|} v(|x_o|,\zeta+|y_o|)) \tag{12}$$

$$= T_k^{\xi,\zeta,|x_o|,|y_o|} \overline{I}(v(|x_o|,\zeta+|y_o|)) \tag{13}$$

$$= T_k^{\xi,\zeta,|x_o|,|y_o|} v(\xi+|x_o|,\zeta+|y_o|)$$

$$= S_x^{\xi+|x_o|} S_y^{\zeta+|y_o|} T_k^{x,y} v(x,y)$$

$$= v_1(\xi,\zeta)$$

$v_2(\xi,\zeta)$ is an analytic function in $|x_o|$ and $|y_o|$ as well as $\xi$ and $\zeta$, therefore it is meaningful to truncate $v_2(\xi,\zeta)$ with respect to $\xi$, $\zeta$, $|x_o|$, and $|y_o|$. In going from (12) to (13) the truncation of $v(|x_o|,\zeta+|y_o|)$ is dropped. Since only the coefficients of v of order greater than k are disturbed in the initial conditions of the $\overline{I}$ problem in (12), only the coefficients greater than k of the solution of the $\overline{I}$ problem are disturbed. Since the solution is truncated, the original truncation has no effect. Since

$$T_k^{\xi,\zeta,|x_o|,|y_o|} v_2(\xi,\zeta) = v_1(\xi,\zeta)$$

and $v_2(\xi,\zeta)$ has only positive coefficients considered

as a function of $\xi$, $\zeta$, $|x_o|$, and $|y_o|$, it follows that $v_2(\xi,\zeta) \supseteq v_1(\xi,\zeta)$. $v_1(\xi,\zeta)$ has only positive coefficients, therefore it dominates zero.

Putting Lemmas 1 and 2 together:

$$V(\xi,\zeta) - v_1(\xi,\zeta) \supseteq V(\xi,\zeta) - v_2(\xi,\zeta) \supseteq U(\xi,\zeta) - u_2(\xi,\zeta).$$

Therefore

$$S_\xi^{|\xi_o|} S_\zeta^{|\zeta_o|} T_{[sk]}^{\xi,\zeta} (V(\xi,\zeta)-v_1(\xi,\zeta)) \tag{14}$$

is greater than or equal to

$$S_\xi^{|\xi_o|} S_\zeta^{|\zeta_o|} T_{[sk]}^{\xi,\zeta} (V(\xi,\zeta) - v_2(\xi,\zeta)) \tag{15}$$

is greater than or equal to

$$\left| S_\xi^{\xi_o} S_\zeta^{\zeta_o} T_{[sk]}^{\xi,\zeta} (U(\xi,\zeta) - u_2(\xi,\zeta)) \right| \tag{16}$$

(14) is the error of the second kind when the first method of continuation is applied to the dominating problem. (15) is the error of the second kind when the second method of continuation is applied to the dominating problem. (16) is the absolute value of the error of the second kind when the second method of continuation is applied to the original problem. Since $(|x_o|,|y_o|)$ is in the domain of convergence of $v(x,y)$ expanded about the origin and s is appropriately chosen, (14) converges to zero by Section II. Therefore (16) converges to zero, which completes the proof.

With almost no changes in the statement and proof of convergence, the second method can be expanded to include any system of quasi-linear partial differential equations in any number of unknowns. It should be pointed out that this also includes ordinary differential equations. The domain of convergence of the power series of one unknown about some point must be altered to mean the largest open set in which all of the power series of the unknowns expanded about this point converge absolutely. Also, the definition of A and B must be appropriately altered.

Since (14) is greater than or equal to (15) one might suspect that the second method of continuation is better than the first. This actually proves to be true in some simple examples as well as in the computations done on the detached shock problem. These results are described in the next section.

## 4. Results of the Test Calculations.

Description of the Detached Shock Problem. Let $z,r$ be cylindrical coordinates. It was assumed that the flow is symmetric about the z-axis, and that the shock is given by $z = H(r)$ where $H(r)$ is an analytic function. In the region ahead of the shock, given by $z < H(r)$, the flow quantities are constant and the flow is parallel to the z-axis. The shock was assumed to be a hyperboloid of revolution, the gas was assumed to be polytropic with a $\gamma$ of 1.4, and M (the mach number) was taken to be 12. Since the publishing of the earlier report (6), it has been pointed out by G. B. Whitham and others that, with such a large Mach number, $\gamma = 1.4$ is not realistic, owing to dissociation of the gas. In order to be realistic $\gamma = 1.2$ would be a better approximation for air. However, since the primary interest in the computation was theoretical, $\gamma = 1.4$ was retained so that the present results could be compared with those from the Univac.

In order to apply the methods described it is necessary to introduce a curvilinear coordinate system which has the property that one coordinate is constant along the shock. This is the only property that the coordinates must satisfy other than that the functions involved in defining the coordinate system must be

analytic. In order to achieve this the coordinates x and y were introduced:

$$x = z - H(r) \qquad\qquad z = x + h(y)$$
$$\text{or}$$
$$y = r - r_o \qquad\qquad r = y + r_o$$

where $r_o$ is a constant and $h(y) = H(y+r_o)$. These particular coordinates have the property y is a radial coordinate measured from the point $r = r_o$ (not from the axis) and x is a coordinate measured parallel to the axis from the shock (not from a fixed plane). The flow quantities were so scaled that the radius of curvature at the tip of the shock was 1. This guarantees that the tip of the body is located at $z = .1$, $r = 0$ approximately.

Let $\rho$ be the density, p the pressure, u the component of velocity parallel to the axis of symmetry, and v the component perpendicular to the axis. Using the Rankine-Hugoniot shock conditions the flow quantities (u, v, $\rho$, and p) can be determined as functions of y when $x = 0$. The differential equations are easily derived from the law of conservation of mass, the law of conservation of momentum, the assumption that changes are adiabatic (that is, entropy is constant along stream-lines), and the equation of state for a polytropic gas.

Most of the computations were directed toward finding the location of the body. The surface of the body can be

defined in terms of the stream function, $\psi$. $\psi$ may be determined in terms of the other flow quantities by means of quadratures. In this problem the streamline $\psi = 0$ presumably consists of two parts: the z-axis and a curve intersecting the z-axis at a positive value of z. The second part may be taken as defining the surface of the body.

$r_o$ was chosen and the flow quantities ($u$, $v$, $\rho$, and $p$) were approximated by truncated power series in x and y. (In terms of the z,r coordinate system the expansions were made about the point $(H(r_o), r_o)$; changing $r_o$ allows one to investigate different parts of the surface of the body.) Using the notation of Section III, $(x_o, y_o)$ was chosen: and, then, using either the first or the second method of continuation, the flow quantities were approximated by truncated power series in $\xi$ and $\zeta$ where $\xi = x - x_o$ and $\zeta = y - y_o$. Both the truncated series in x and y and in $\xi$ and $\zeta$ were computed to a fixed order k. In most of the computations k was taken to be 19 (the largest possible k due to machine restrictions). The truncated power series in $\xi$ and $\zeta$ of the flow quantities were then further truncated so that all of the terms of order greater than [sk] were discarded. Using these approximations to $u$, $v$, $\rho$, and $p$, an approximation to $\psi(\xi, \zeta)$ was computed; call this

$\psi_1(\xi,\zeta)$. $\zeta_o$ was chosen and finally $\xi_o$, the solution to $\psi_1(\xi,\zeta_o) = 0$, was computed. In the original z,r coordinate system, then,

$$(x_o+\xi_o+h(y_o+\zeta_o), y_o+\zeta_o+r_o)$$

is an approximation to a point on the surface of the body.

Results Using the Second Method. In order to compute a point on the body surface, six constants ($k$, $r_o$, $x_o$, $y_o$, [sk], and $\zeta_o$) must be chosen. This number was too great to allow a thorough investigation of the effect on the accuracy of varying these constants. In all of the results discussed, $k = 19$. The constants for which the best results were obtained are indicated schematically in Figure 3 page 39. For example, with the choice of constants $r_o = .4$, $x_o = .03$, $y_o = 0$, $\zeta_o = -.1$, first the truncated power series in x and y about the point A were obtained, then truncated power series in $\xi$ and $\zeta$ about the point B, and then the body point C. $\zeta_o$ was always chosen a multiple of .05, and $y_o$ was always taken to be zero. The best results were obtained by the following combinations of $x_o$ and $r_o$: $r_o = .2$, $x_o = .03$; $r_o = .4$, $x_o = .03$; $r_o = .5$, $x_o = .04$; $r_o = .8$, $x_o = .08$. Although few cases were tried other than these already listed, it seemed clear that $(x_o,y_o)$

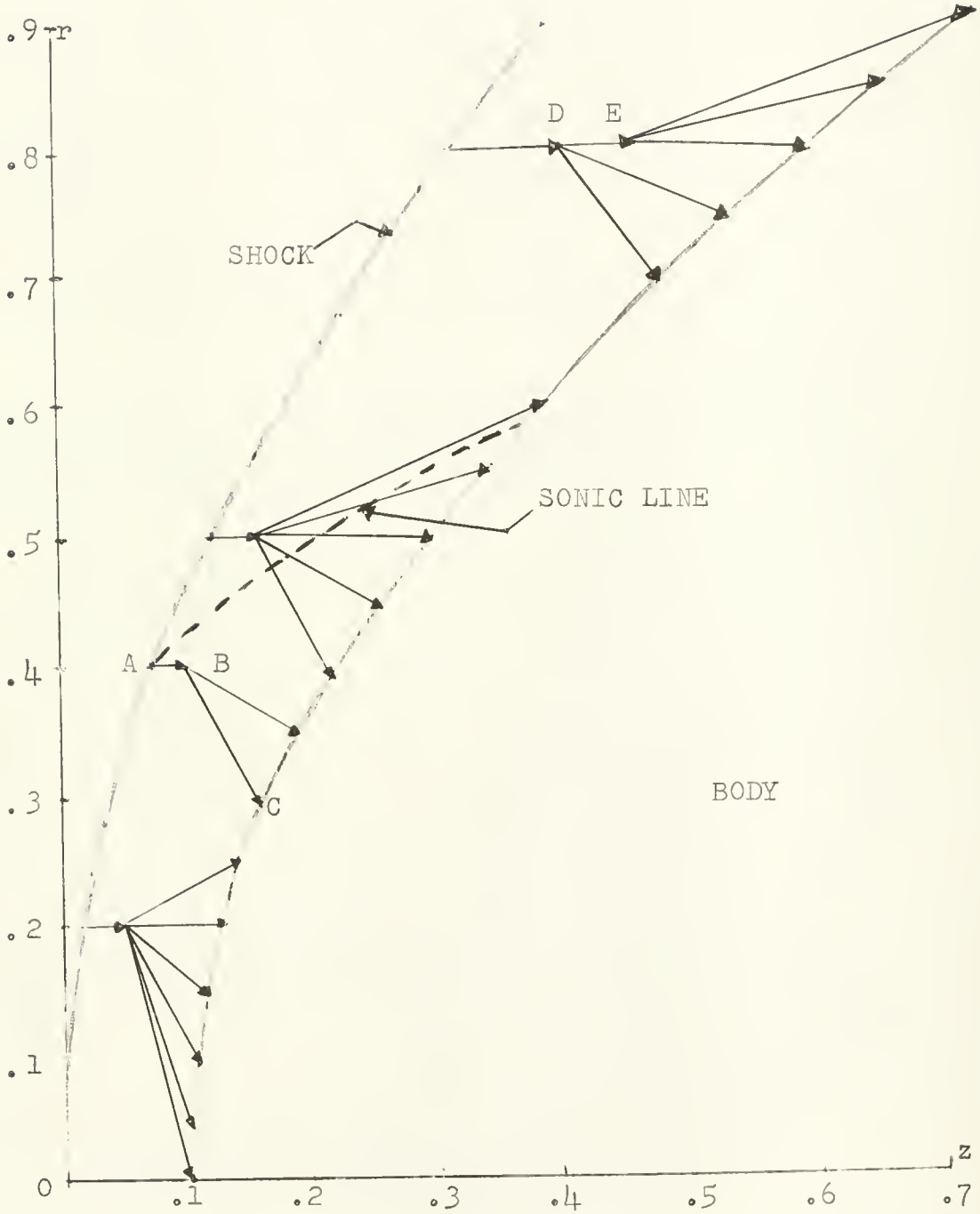


Figure 3

must be taken closer to the shock than to the body. The point E in Figure 3 requires some explanation. The truncated power series about this point was not obtained directly from the point on the shock, but was obtained through the use of the second method of continuation from the truncated power series about D. The results were never improved by taking [sk] less than 19.

In choosing the best result, three checks were available:

1) Bernoulli's law states that $(u^2+v^2)/2 + \gamma p/((\gamma-1)\rho)$ should be constant (208.6 for this particular problem) throughout the flow.

2) Consider some point in the flow where the pressure and density are p and $\rho$. Let $p_1$ and $\rho_1$ be the pressure and density of the flow just after the streamline which goes through this point crosses the shock. Since entropy is constant along streamlines, $p/p_1$ should be equal to $(\rho/\rho_1)^\gamma$. The details of computing $p_1$ and $\rho_1$ are described in Appendix A. This check proves to be a little too sensitive, probably due to the great number of approximations involved.

3) The flow quantities at a point may be computed with different choices of $r_o$, $x_o$, etc.; these of course should agree.

To illustrate these checks the following example is given:

| | r | z | Ber. | $(\rho/\rho_1)^\gamma$ | $p/p_1$ |
|---|---|---|---|---|---|
| (1) | .05 | .10328060 | 208.60000 | 1.1015690 | 1.1015690 |
| (2) | .05 | .10328060 | 208.60001 | 1.1015696 | 1.1015695 |

The entrances in row (1) were computed with $r_o = .2$, $x_o = .03$, and $\zeta_o = -.15$; the entrances in row (2) with $r_o = 0$, $x_o = .03$, and $\zeta_o = .05$. $y_o = 0$ in both cases. It is not unreasonable to assume that the figures given in column z have eight digits of accuracy. Of course, there is always the possibility that these figures are deceiving. But this could always be true for any method for which there is no exact check on the answers. In the present case it seems very unlikely.

Table 1 on page 62 gives the location of the surface of the body in the z,r coordinate system for various values of r and also the flow quantities. Only the digits which were felt to be accurate are listed.

It should be pointed out that the power series which were computed bear out the conjecture that some method of continuation must be used in order to compute the surface of the body. When $r_o = 0$, the first 20 coefficients of $u(x,0)$, $p(x,0)$, and $\rho(x,0)$ behave approximately like the first 20 coefficients of $1/(.08+x)$. Therefore the body would seem to be outside

the domain of convergence of the power series of $u(x,y)$. This same property seems to be true for all the truncated power series expanded about points on the shock surface.

Results of the First Method. Since the results using the second method seem to be accurate to eight digits, these results were used as a guage for checking the accuracy of the results using the first method. All of the computations using the first method used $k = 19$ and $r_o = .2$. Various choices for $x_o$ and $y_o$ were used: $x_o = .03$, $y_o = -.1$; $x_o = .03$, $y_o = 0$; $x_o = .05$, $y_o = 0$; $x_o = .03$, $y_o = .1$. In all cases which were computed, $[sk] = 8$, $x_o = .03$, and $y_o = .1$ gave the most accurate results.

| | 1st method | 2nd method |
|---|---|---|
| r | z | z |
| .1 | .10852240 | .10852384 |
| .2 | .12960302 | .12960099 |
| .3 | .16509492 | .16509964 |

The first method, then, gives approximately 6 digits of accuracy.

The Choice of $x_o$ and s in the First Method. The first method does not involve in any essential way the use of the differential equations. It can be thought of as a method for "analytically" continuing a power series. If the method were applied to an analytic function of one variable, in most cases $x_1$ and k would be given. The questions is, then, how to choose $x_o$ and s so as to minimize the error of the approximation. For simplicity assume that the analytic function is given as a power series about the origin, that the radius of convergence of this power series is 1, and that the only singularity that the function possesses is at $x = a$. (From this it follows that $|a| = 1$.) In order to gain values for $x_o$ and s two important assumptions are made. First, it is assumed that there is no systematic cancellation between the errors of the first and second kind as described in Section II. In that section an upper bound was derived for $|E_2|$ (the error of the second kind). Although it is not clear at first glance, all of the estimates used in deriving this bound are extremely tight. The second assumption is that the error of the second kind actually behaves like this upper bound. These two assumptions may well always hold. Under these assumptions it must be true that the best choice for $x_o$ and s must be such as to make the exponential part of the

error of the second kind equal to the exponential part of the error of the first kind:

$$|x_o| \left|\frac{x_o - x_1}{x_o}\right|^s \frac{1}{s^s(1-s)^{1-s}} = \left|\frac{x_1 - x_o}{a - x_o}\right| \tag{17}$$

The best choice for $x_o$ and s would then be the value of $x_o$ and of s that minimizes the right hand side (or the left hand side) of (17). For any $\varepsilon > 0$, the error in the approximation for large k would be less than

$$M\left(\left|\frac{x_1 - x_o}{a - x_o}\right| + \varepsilon\right)^k$$

where M is some constant which depends on $\varepsilon$ and f(x). Even if the two assumptions were dropped, the previous statement would still be true, but it might be possible to choose $x_o$ and s so as to obtain an even smaller bound. The values, $x_o$ and s, which minimize one side of (17) are given in Figure 4 on the next page as functions of $x_1$ for various values of a.

$$E = \left|\frac{x_1 - x_o}{a - x_o}\right|^s$$

is also listed as a function of $x_1$. In Figure 4 the graphs numbered 1 correspond to a = -1, those numbered 2 correspond to a = $\pm i$, and those numbered 3 correspond to a = $1/2 \pm \sqrt{3}/2\, i$. In Tables 2 and 3 on pages 63 and 65, these values of $x_o$ and s are used to compute the

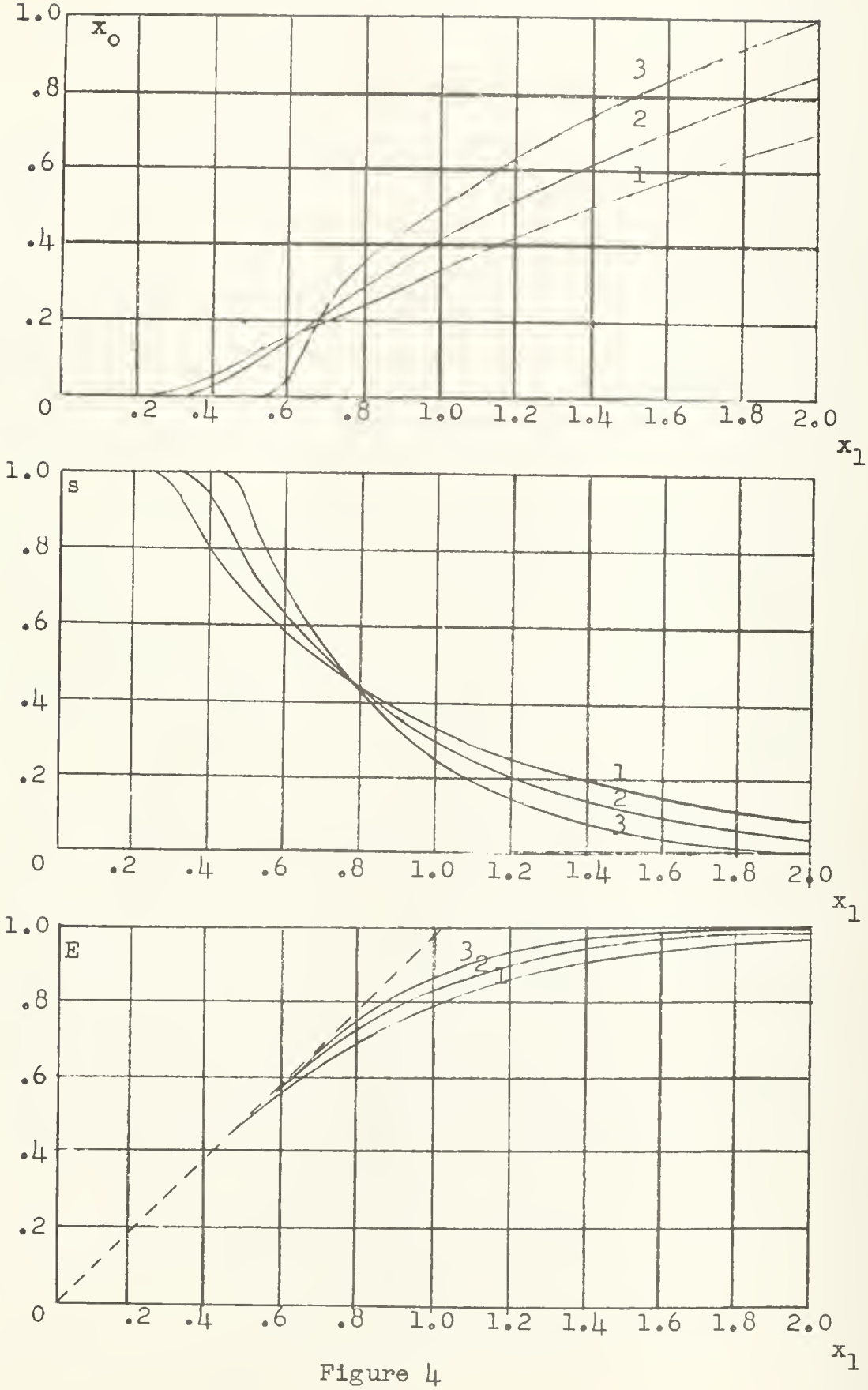


Figure 4

values of various elementary functions at $x_1 = .8$ and $x_1 = 1.2$. The machine also computed the best choice for [sk] when $f(x)$, $x_o$, $x_1$, and k were given. This best choice differed so little from the [sk] gained from the fixed s computed through (17) that it is not listed. This supports the belief that the two assumptions used to derive (17) were very reasonable. The error was at times a factor of 10 better when the best choice for [sk] was used, but this was never consistent. As can be seen from the tables, the error seems to be affected as much by large coefficients in the original power series (e.g. $(1/(1+x)^3)$ as by a near singularity (e.g. $1/(1+x^3)$).

<u>Choice of $x_o$ and s in the Second Method</u>. How to make the best choice for $x_o$ and s in the second method is a much more complicated question. It seems to depend not only on the solution but also on the differential equations involved. In the detached shock computations it seemed clear that the best choice for s was s = 1. It is easy to find examples in which s = 1, if not the best choice for s, is the only reasonable choice for s given a particular k, $x_o$, etc.. (The best choice for s can be unreasonable in the following sense: the use of the best [sk] in the truncation produces the smallest error, but the use of [sk]+1 produces a very large error.) One of these examples is as follows:

$$\frac{du}{dx} = -u^2 \quad \text{where} \quad u(0) = 1.$$

The solution is $u = 1/(1+x)$. If the second method of continuation is used for $x_1 = 1$, $x_o = \sqrt{2} - 1$ and $s = 1$ are the most reasonable choices. The error is then asymptotic to $(-1)^k(1-\frac{\sqrt{2}}{4})(\sqrt{2}-1)^{k+1}$. If the first method is used then $x_o = 1/3$ and $s = 1/3$ are the best choices according to the last section and the error is bounded by $M(\frac{1}{2^{1/3}} + \varepsilon)^k$ or $M(.794+\varepsilon)^k$ for $\varepsilon > 0$ and M large enough. This bound was confirmed by actual test computations. Hence, the second method is much superior to the first in this example.

Since in both the previous example and in the detached shock computations the best choice for s using the second method of continuation is apparently $s = 1$, it should be pointed out that there exist cases where s cannot be so chosen. If the partial differential equations are linear with constant coefficients then the first and second methods of continuation are identical. Therefore if there is a singularity in the initial conditions, there exist points for which neither method will converge unless s is taken less than one.

Concluding Remarks. Even though the methods described give great accuracy, in actual practice this is balanced by the fact that a large amount of computing time is required. In order to compute the truncated power series of the flow quantities from the differential equations and the initial conditions when $k = 19$, 20 minutes is required on the IBM 704. If all that is required is the stand-off distance of the shock from the body, using either method, the time required for the rest of the computation is negligible. Therefore the first method requires approximately 20 minutes and the second method 40 minutes. If the arithmetic (see Appendix B) were simplified and the code rewritten, the time could be reduced by at most a factor of 2.

Other methods are known for solving the detached shock problem. P. R. Garabedian and H. M. Leiberstein, and M. D. Van Dyke, for example, have described successful finite-difference methods, see [11] and [8]. Van Dyke's method gives the stand-off distance to two figures of accuracy (which is all that is required for practical purposes) in 5 minutes on a much slower machine. Even if the two methods of continuation were streamlined as much as possible, it seems doubtful if they could ever compete with Van Dyke's computing time.

It is not remarkable that the first method is

insensitive to whether the equations involved are elliptic or hyperbolic, but the second method seems insensitive also. In other methods, such as Van Dyke's special consideration must be given to the sonic line. With the present methods no effect whatever was noticed when crossing the sonic line.

Even though convergence has been shown, there are many questions which need to be answered: Under what conditions is the second method preferable to the first? Are the restrictions on s in Section II the weakest possible under the given assumptions? (This seems very likely.) Is the use of the dominating problem in Section III ever necessary, and for what conditions may $s = 1$? May the second method be applied repeatedly? How should $x_o, y_o$ and s be chosen in the second method so as to assure the maximum rate of convergence?

Changing one's point of view might also yield interesting results. For example, is it possible to obtain a useful statement of the first method in terms of information theory? The first method could be thought of as a means for recovering an encoded message (the power series of a function) when there is noise (truncation). This is made plausible by the fact that $\log(g(z))$ contains a term suggestive of entropy or information content.

## Appendix A. Details of the Test Problem.

The Equations. Let $z, r$ be cylindrical coordinates, and let the shock be given by $z = H(r)$. It is convenient to take the units of length, density, and pressure (these are dimensionly independent) to be the radius of curvature of the shock at the axis, the density ahead, and the pressure ahead. Then the conditions ahead may be written as

$$p = 1$$

$$\rho = 1$$

$$u = U$$

$$v = 0$$

Here $u$ and $v$ are axial and radial components of fluid velocity and $U$ is a constant related to the Mach number by the equation

$$U = M\sqrt{\gamma}$$

where $\gamma$ is the constant of the polytropic gas.

It was assumed that $H(r)$ satisfied the equation

$$H(r) = \sqrt{\alpha^2 + \alpha r^2} - \alpha$$

where $\alpha = M^2 - 1$. This means the shock was assumed to be a hyperboloid of revolution, with a radius of curvature equal to 1. The $\alpha$ was so chosen that the shock approximates a sonic disturbance as $r$ becomes large. This condition would be satisfied by every shock created by the motion of a finite blunt body.

Introduce the coordinates $x$ and $y$ defined by:

$$x = z - H(r) \qquad\qquad z = x + h(y)$$

$$y = r - r_o \qquad \text{or} \qquad r = y + r_o$$

where $r_o$ is a constant and $h(y) = H(y+r_o)$.

Since the flow quantities in front of the shock and the shock shape are assumed, the flow quantities immediately behind the shock are determined as functions of y by the Rankine-Hugoniot shock conditions for a polytropic gas.

$$\rho = \frac{\theta U^2}{(\theta+1)(1+h'^2)+U^2} \tag{1}$$

$$p = \frac{\theta\rho - 1}{\theta - \rho}$$

$$u = U - \frac{(\theta-1)U}{\theta(1+h'^2)}$$

$$v = (U-u)h'$$

where $\theta = \frac{\gamma+1}{\gamma-1}$, $h'$ stands for $h'(y)$, and u, v, $\rho$, and p stand for $u(0,y)$, $v(0,y)$, $\rho(0,y)$, and $p(0,y)$. Therefore the expansion of u, v, $\rho$, and p in powers of y can be obtained by these equations from the expansion of $h(y)$.

The flow quantities, u, v, $\rho$, p, and S (entropy), satisfy the following equations:

$$(r\rho u)_z + (r\rho v)_r = 0 \tag{2}$$

$$uu_z + vu_r + p_z/\rho = 0 \tag{3}$$

$$uv_z + vv_r + p_r/\rho = 0 \tag{4}$$

$$uS_z + vS_r = 0 \tag{5}$$

$$p = A(S)\,\rho^{\gamma} \tag{6}$$

where A(S) is some analytic function of S. (2) is the law of conservation of mass, (3) and (4) are the laws of conservation of momentum, (5) is the equation which results from the assumption that the changes are adiabatic, and (6) is the equation of state for a polytropic gas. The entropy can be eliminated between equations (5) and (6).

These equations may be rewritten in the x,y coordinate system in the following form:

$$\frac{\partial p}{\partial x} = \frac{\rho(u-vh')(v\frac{\partial p}{\partial y} + \frac{\gamma vp}{y+r_o}) + \gamma p(\rho u\frac{\partial v}{\partial y} + h'\frac{\partial p}{\partial y} - \rho v\frac{\partial u}{\partial y})}{\gamma p(1+h'^2)-\rho(u-vh')^2} \tag{7}$$

$$\frac{\partial u}{\partial x} = -\frac{\frac{\partial p}{\partial x} + \rho v\frac{\partial u}{\partial y}}{\rho(u - vh')} \tag{8}$$

$$\frac{\partial v}{\partial x} = -\frac{\frac{\partial p}{\partial y} - h'\frac{\partial p}{\partial x} + \rho v\frac{\partial v}{\partial y}}{\rho(u-vh')} \tag{9}$$

$$\frac{\partial \rho}{\partial x} = \frac{\rho(u-vh')\frac{\partial p}{\partial x} + \rho v\frac{\partial p}{\partial y} - \gamma pv\frac{\partial \rho}{\partial y}}{\gamma p(u-vh')} \tag{10}$$

The only term which could possibly not be analytic is $\gamma vp/(y+r_o)$ in equation (7). If this were singular when $r_o = 0$, it would be impossible to compute the flow quantities along the axis. But, presumably, $v = 0$ along the axis; hence $\gamma vp/(y+r_o)$ should be analytic and give no trouble.

In terms of $\rho(x,y)$ and $v(x,y)$, the stream function

$\psi$ is given, except for a factor of $2\pi$, by the equation

$$\psi = \psi(x,y) = \frac{1}{2} U (y+r_o)^2 - (y+r_o) \int_o^x \rho(x',y)v(x',y)dx'.$$

In actual computation the truncated power series for $\int_o^x \rho(x',y)v(x',y)dx'$ is obtained by multiplying the truncated power series for $\rho$ and $v$ together and integrating term by term.

<u>Computation of the Coefficients</u>. In both methods of continuation it is necessary to compute the coefficients of the truncated power series of the unknowns from the differential equations and the initial conditions. In order to facilitate the discussion of the computation of the coefficients it is useful to introduce the following notation: let $f_{i,j}$ be the coefficient of the $x^i y^j$ term in the power series expansion of $f(x,y)$; that is, $f(x,y) = \sum f_{i,j} x^i y^j$ where the sum is taken over all non-negative i and j. In the same way, let $h_j$ be such that $h(y) = \sum h_j y^j$.

To start the computation, the machine reads in the four constants $r_o$, k, M, and $\gamma$. By an iteration formula, the machine then computes $h_j'$ for all non-negative j less than or equal to k. The machine then uses these coefficients together with equations (1) to compute $u_{o,j}$, $v_{o,j}$, $\rho_{o,j}$, and $p_{o,j}$ for all j less than or equal to k, and places them in appropriate storage locations. This

ends the first phase of the computation. The computation of the coefficients when $i = 0$ is similar to the computation when $i > 0$ and will not be discussed further.

In order to compute the coefficients when $i > 0$, the machine contains subroutines which can form the sum, difference, product, and quotient of two power series. To be explicit, the subroutine is instructed on entry as to which coefficient is to be computed, the location of the coefficients of the two power series to be operated on, and the location where the coefficients of the power series of the sum, difference, etc., are to be stored. There are also subroutines to differentiate and integrate power series. The subroutines are based on the following simple identities (it is assumed that the coefficients of $f(x,y)$ and $g(x,y)$ are known):

If $f(x,y)g(x,y) = h(x,y)$ then

$$h_{I,J} = \sum_{i=0}^{I} \sum_{j=0}^{J} f_{i,j}\, g_{I-i,J-j} \,. \tag{11}$$

If $f(x,y)/g(x,y) = h(x,y)$ then $h_{I,J} =$

$$\frac{1}{g_{o,o}}\left(f_{I,J} - \sum_{i=0}^{I-1} \sum_{j=0}^{J-1} h_{i,j}\, g_{I-i,J-j} - \sum_{j=0}^{J-1} h_{I,j} g_{o,J-j}\right). \tag{12}$$

If $\frac{\partial f(x,y)}{\partial y} = h(x,y)$ then

$$h_{I,J} = (J+1)\, f_{I,J+1} \,. \tag{13}$$

If $f(x,y) = \frac{\partial h(x,y)}{\partial x}$ and $J \geq 1$ then

$$h_{I,J} = \frac{1}{J} f_{I,J-1} \cdot \tag{14}$$

In (12) if either $I = 0$ or $J = 0$ the appropriate sum should be dropped.

With the aid of the subroutines the coefficients of u, v, $\rho$, and p for $i > 0$ are easily computed. Assume that $u_{i,j}$, $v_{i,j}$, $\rho_{i,j}$, and $p_{i,j}$ are known for all $i+j \leq k$, $i \leq I$. From these $(\frac{\partial u}{\partial y})_{i,j}$, $(\frac{\partial v}{\partial y})_{i,j}$, etc. are computed for $i+j \leq k-1$, $i \leq I$ (differentiation lowers the order). Every power series on the right of equation (7) is known for $i+j \leq k-1$, $i \leq I$, hence $(\frac{\partial p}{\partial x})_{i,j}$ can be computed using the subroutines for $i+j \leq k-1$, $i \leq I$. The integration subroutine gives $p_{i,j}$ for $i+j \leq k$, $i = I+1$ (integration raises the order). In a similar manner using equations (8) to (10), the coefficients of u, v, and $\rho$ for $i+j \leq k$, $i = I+1$ are computed. Hence the machine essentially does an induction on I until all of the coefficients of order less than or equal to k are computed. Actually the machine stores the coefficients of 23 different power series such as $v \frac{\partial \rho}{\partial y}$, $\rho(u-vh')$, etc., in order to avoid the duplication of computations.

Certain restrictions are necessary on the functions which occur in the differential equations if actual computations are to be made feasible. Essentially, the

functions of the dependent variables must be easily expressible as power series in the independent variables when the power series for the dependent variables are known (i.e. simple recursion formulas involving the coefficients of the power series of the dependent variables must give the coefficients of the power series of the functions). The previous section has just shown that functions which are made up of sums, differences, products, and quotients of the variables and their derivatives are of this type. It is possible for the functions to contain such terms as sin(x) or even $\sqrt{u}$ ; but sin(v), for example, would require a prohibitive amount of machine computation as well as storage space.

Even though the machine used was the IBM 704, the subroutines are essentially those used by the Univac in the earlier computations done by R. D. Richtmyer. The earlier report (6) gives a more detailed account of the subroutines, if more information on the subroutines is desired.

<u>The Entropy Check</u>. Let $\rho$, p, and $\psi$ be the density, pressure, and the stream function at some point in the flow. The streamline through this point presumably intersects the shock. Let $(0,\bar{y})$ be this point of intersection in the x,y coordinate system and let $\rho_1$, $p_1$ be the density and pressure on the downstream side

of the intersection. Since the stream function is constant along streamlines

$$\psi = \frac{1}{2} U(\bar{y}+r_o)^2$$

or

$$\bar{y} = \sqrt{\frac{2\psi}{U}} - r_o \ .$$

$\bar{y}$ substituted in equations (1) gives $\rho_1$ and $p_1$. Therefore, it is possible to determine $\rho_1$ and $p_1$ from the knowledge of $\psi$.

Since entropy is constant along streamlines it follows from (6) that

$$\frac{p}{p_1} = \left(\frac{\rho}{\rho_1}\right)^\gamma \ .$$

If $\rho$, $p$, and $\psi$ are approximated, it is possible therefore to find an approximation to $(p/p_1)$ and $(\rho/\rho_1)^\gamma$. Comparing these last two approximations will give an indication of the accuracy of the original approximations.

<u>Input, Output, Etc.</u> To start the computation $r_o$, $k$, $\gamma$ and $M$ are fed into the machine on cards. The main code then computes the coefficients of the power series in $x$ and $y$ of $u$, $v$, $\rho$, $p$ and $\rho v$ (used to compute $\psi$) and punches the results out on cards. If the second method of continuation is being used, this result is fed back into the machine with the constants $x_o$, $y_o$; and then, the coefficients of the power series in $\xi$, $\zeta$ of $u$, $v$, $\rho$, $p$ and $\rho v$ are punched out.

With the first method, the coefficients of the

power series in x and y are fed into the machine with the output code plus the constants $x_o$ and $y_o$. The machine computes the power series in $\xi$ and $\zeta$ and reads in a control card. With the second method, the coefficients of the power series in $\xi$ and $\zeta$ are fed into the machine with the output code and a control card.

This control card for both methods contains [sk], $\xi_1$, $\delta\xi$, m, $\zeta_1$, $\delta\zeta$, and n. These last six constants are to instruct the machine to print out on the tabulator the flow quantities at the points $(\xi_1+i\delta\xi,\ \zeta_1+j\delta\zeta)$ where i =0 to m and j = 0 to n. First the machine prints all the flow quantities at the points $(\xi_1+i\delta\xi,\zeta_1)$ i = 0 to m, and then proceeds to search for the point on the surface of the body with $\zeta = \zeta_1$. In order to do this, it is necessary to solve the equation $\psi(\xi,\zeta_1)=0$. This is done by Newton's method for finding the roots of an equation. The flow quantities are then printed out for this point on the body. Next the machine prints out the flow quantities for the points $(\xi_1+i\delta\xi,\zeta_1+\delta\zeta)$ i = 0 to m; etc.

The following flow quantities are printed out: u, v, $\rho$, p, $\psi$, $(\rho/\rho_1)^\gamma$, $(p/p_1)$, and $(u^2+v^2)$. $(u^2+v^2)$ is printed out in order to have a simple indication as to whether the flow is subsonic or supersonic. Let $\hat{q}$ be the Bernoulli number and $c^*$ be the critical speed.

These are related by the equation:

$$c^{*2} = \frac{\gamma-1}{\gamma+1}\,\hat{q}^2$$

(For $\gamma = 1.4$ and $M = 12$, $c^{*2} = 34.766666$.)

If $(u^2+v^2) > c^{*2}$ the flow is supersonic.

If $(u^2+v^2) < c^{*2}$ the flow is subsonic.

## Appendix B. The Arithmetic Used in the Test.

When the power series method for solving the detached shock problem was first coded for the Univac, a straight forward single-precision floating-decimal arithmetic was used. This arithmetic proved unsuccessful. Some of the high order coefficients were so inaccurate that it was necessary to discard these coefficients before the power series could be evaluated. In order to do this it was necessary to distinguish between the accurate and the inaccurate coefficients. This was accomplished by recoding the detached shock problem with single-precision floating-decimal subroutines with significance monitoring. Each numerical quantity was represented by not only a fraction and an exponent but also a significance index. This significance index was an integer which indicated roughly how many digits in the fraction were accurate. This monitoring scheme proved very successful. (See [6] for a description of these subroutines.)

When the problem was recoded for the IBM 704 it was decided to use double-precision and to retain accuracy monitoring but to use an alternative procedure suggested by N. C. Metropolis. Ordinarily the fractional part of a floating-point number on the 704 (a binary machine) has a leading one. Instead of keeping a

significance index, enough leading zeros are supplied to the fraction so that all the remaining bits are significant. Because leading zeros are allowed in the fraction, this arithmetic is called unnormalized. This monitoring scheme has the advantage that no bits are wasted in storing the significance index.

Accuracy monitoring for this unnormalized floating-point arithmetic is done as follows: Ordinarly, when two floating-point numbers are added or subtracted, the last step in the operation consists in normalizing the answer (that is, the fraction of the answer is shifted left and the exponent is reduced until there is a leading one in the fraction). In the present arithmetic this last step is suppressed. In multiplication and division the number of leading zeros in the fraction of the answer is exactly the same as the number of leading zeros in the least significant of the two numbers entering into the operation.

To check the monitoring, the arithmetic was used to compute sin(x) by its power series for large values of x. The results are listed in Table 4, page 67. Column A lists the number of significant digits (i.e. not counting the leading zeros) in the fraction out of the possible 54. Column B lists the number of accurate bits in the fraction. Column C lists the number of

accurate bits in the fraction when normalized arithmetic is used. Considering that the monitoring scheme does not take into account the accumulation of error due to round-off, the results are surprisingly good.

Details of this arithmetic and the code itself may be obtained by writing to the SHARE Librarian, 590 Madison Avenue, New York 22, New York, and asking for NU-UDP1.

In all of the computations so far the significance monitoring was indispensable. Even though the coefficients computed on the 704 usually have a singificance of 30 bits or more, no reliance could be placed on the coefficients of large order without the monitoring. Now that more work has been done on the shock problem, it is no doubt possible to eliminate the monitoring. This would mean a great saving in time. Since, in all of the computations, the significance of the coefficients varies inversely with the order, it would seem reasonable to rely on this fact in future computations. k can be varied and the accuracy of the computation can be checked (by means of Bernoulli's law and the entropy check) until a reasonable k is found. If k is too large the approximation is corrupted by round-off error; if k is too small the approximation is corrupted by truncation error.

Table 1

The Location of the Body

Results Using the Second Method of Continuation

| r | z | u | v | ρ | p |
|---|---|---|---|---|---|
| 0 | .10153518 | 0 | 0 | 6.237223 | 185.8692 |
| .05 | .10328060 | .039276 | .5623781 | 6.213493 | 184.8800 |
| .1 | .10852384 | .155788 | 1.113109 | 6.143220 | 181.9593 |
| .15 | .11728574 | .345687 | 1.641201 | 6.029069 | 177.2434 |
| .2 | .1296009 | .602890 | 2.136897 | 5.87629 | 170.9442 |
| .25 | .1455182 | .919511 | 2.59213 | 5.68708 | 163.329 |
| .3 | .16509964 | 1.28661 | 3.000793 | 5.47065 | 154.694 |
| .35 | .1884216 | 1.6943 | 3.35889 | 5.2325 | 145.35 |
| .4 | .21613294 | 2.132750 | 3.664414 | 4.979048 | 135.5891 |
| .45 | .24666008 | 2.592515 | 3.917230 | 4.716305 | 125.6789 |
| .5 | .2817980 | 3.06474 | 4.11878 | 4.44981 | 115.849 |
| .55 | .321123 | 3.5400 | 4.2722 | 4.1858 | 106.33 |
| .6 | .36484 | 3.988 | 4.388 | 3.949 | 97.87 |
| .7 | .46570233 | 4.941076 | 4.476248 | 3.4261057 | 80.34169 |
| .75 | .52335352 | 5.382131 | 4.474349 | 3.194297 | 72.83585 |
| .8 | .5860619 | 5.80537 | 4.44483 | 2.97533 | 65.9419 |
| .85 | .654066 | 6.2057 | 4.3932 | 2.7737 | 59.739 |
| .9 | .72829 | 6.528 | 4.349 | 2.667 | 55.88 |

Table 2

Results Using the First Method of Continuation on Elementary Functions with $x_1 = 1.2$

Description

The results are tabulated (on the next page) when the first method of continuation is used to compute the value of the functions $1/(1+x)$, $1/(1+x)^3$, $1/(1+x^2)$, and $1/(1+x^3)$ at $x_1 = 1.2$ when only the first k terms of the power series about the origin of these functions are used. In accordance with the discussion on page 42, $x_o$ and s were chosen in the following way:

$x_o = .41392785$ and $s = .255$ when computing $1/(1+x)$ and $1/(1+x)^3$.

$x_o = .51720723$ and $s = .202$ when computing $1/(1+x^2)$.

$x_o = .62752519$ and $s = .147$ when computing $1/(1+x^3)$.

The columns headed [sk] are actually [sk+1/2]. The error listed is the approximation minus the true value. 8. -3 should be read 8 times $10^{-3}$.

Results Using the First Method of Continuation

on Elementary Functions with $x_1 = 1.2$

The description of the table is on the previous page.

| | $1/(1+x)$ | | $1/(1+x)^3$ | | $1/(1+x^2)$ | | $1/(1+x^3)$ | |
|---|---|---|---|---|---|---|---|---|
| k | [sk] | error | [sk] | error | [sk] | error | [sk] | error |
| 0 | 0 | 5. -1 | 0 | 9. -1 | 0 | 6. -1 | 0 | 6. -1 |
| 5 | 1 | -2. -1 | 1 | -1. 0 | 1 | 7. -2 | 1 | -3. -1 |
| 10 | 3 | 1. -2 | 3 | 4. 0 | 2 | -1. -1 | 1 | -3. -2 |
| 15 | 4 | 1. -2 | 4 | -2. 0 | 3 | 5. -2 | 2 | -2. -1 |
| 20 | 5 | -1. -2 | 5 | 7. -1 | 4 | 3. -3 | 3 | 7. -2 |
| 25 | 6 | 7. -3 | 6 | -2. -1 | 5 | -2. -2 | 4 | 1. -1 |
| 30 | 8 | 4. -3 | 8 | 8. -1 | 6 | 1. -2 | 4 | 7. -2 |
| 35 | 9 | -2. -3 | 9 | -3. -1 | 7 | -2. -3 | 5 | -3. -2 |
| 40 | 10 | 8. -4 | 10 | 1. -1 | 8 | -5. -3 | 6 | -3. -2 |
| 45 | 11 | -4. -4 | 11 | -5. -2 | 9 | 4. -3 | 7 | 8. -3 |
| 50 | 13 | -8. -5 | 13 | 5. -2 | 10 | 1. -4 | 7 | 2. -2 |
| 55 | 14 | 6. -5 | 14 | -2. -2 | 11 | -2. -3 | 8 | 1. -2 |
| 60 | 15 | -3. -5 | 15 | 4. -3 | 12 | 8. -4 | 9 | -6. -3 |
| 65 | 17 | -2. -5 | 17 | -1. -2 | 13 | 3. -4 | 10 | -5. -3 |
| 70 | 18 | 8. -6 | 18 | 5. -3 | 14 | -4. -4 | 10 | -2. -3 |
| 75 | 19 | -4. -6 | 19 | -2. -3 | 15 | 1. -4 | 11 | 3. -3 |
| 80 | 20 | 2. -6 | 20 | 7. -4 | 16 | 6. -5 | 12 | 9. -4 |
| 85 | 22 | 4. -7 | 22 | -7. -4 | 17 | -7. -5 | 12 | 2. -4 |
| 90 | 23 | -3. -7 | 23 | 2. -4 | 18 | 2. -5 | 13 | -1. -3 |
| 95 | 24 | 2. -7 | 24 | -5. -5 | 19 | 1. -5 | 14 | 4. -5 |
| 100 | 26 | 9. -8 | 26 | 1. -4 | 20 | -2. -5 | 15 | 4. -4 |

Table 3

Results Using the First Method of Continuation on Elementary Functions with $x_1 = .8$

Description

The results are tabulated (on the next page) when the first method of continuation is used to compute the value of the functions $1/(1+x)$, $1/(1+x^2)$, and $1/(1+x^3)$ at $x_1 = .8$ when only the first k terms of the power series about the origin of these functions are used. In accordance with the discussion on page 42, $x_o$ and s were chosen in the following way:

$x_o = .24829866$ and $s = .438$ when computing $1/(1+x)$.
$x_o = .29648953$ and $s = .425$ when computing $1/(1+x^2)$.
$x_o = .34043853$ and $s = .418$ when computing $1/(1+x^3)$.

The columns headed [sk] are actually [sk+1/2]. Under each function two errors are listed. The first error column is the approximation minus the true value when the first method of continuation is used. The second error column is the approximation minus the true value when no additional truncation is used (i.e. s = 1). 8. -3 should be read 8 times $10^{-3}$.

Results Using the First Method of Continuation

on Elementary Functions with $x_1 = .8$

The description of the table is on the previous page.

| | $1/(1+x)$ | | | | $1/(1+x^2)$ | | | | $1/(1+x^3)$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| k | [sk] | error | | error | [sk] | error | | error | [sk] | error | | error |
| 0 | 0 | 4. | -1 | 4. -1 | 0 | 4. | -1 | 4. -1 | 0 | 3. | -1 | 3. -1 |
| 5 | 2 | 3. | -2 | -1. -1 | 2 | -6. | -2 | 2. -1 | 2 | -8. | -2 | -2. -1 |
| 10 | 4 | 1. | -2 | 5. -2 | 4 | 2. | -2 | -4. -2 | 4 | -6. | -3 | -5. -2 |
| 15 | 7 | -1. | -3 | -2. -2 | 6 | -3. | -3 | -2. -2 | 6 | 3. | -3 | -1. -2 |
| 20 | 9 | -8. | -5 | 5. -3 | 8 | -8. | -5 | 5. -3 | 8 | -1. | -3 | 6. -3 |
| 25 | 11 | -4. | -5 | -2. -3 | 11 | 1. | -4 | 2. -3 | 10 | 6. | -4 | 2. -3 |
| 30 | 13 | -5. | -6 | 6. -4 | 13 | -6. | -6 | -5. -4 | 13 | 5. | -6 | 4. -4 |
| 35 | 15 | -1. | -6 | -2. -4 | 15 | -6. | -6 | -2. -4 | 15 | -1. | -6 | -2. -4 |
| 40 | 18 | 2. | -7 | 6. -5 | 17 | 2. | -6 | 5. -5 | 17 | -6. | -6 | -6. -5 |
| 45 | 20 | 1. | -8 | -2. -5 | 19 | -3. | -7 | 2. -5 | 19 | 1. | -6 | -1. -5 |
| 50 | 22 | 5. | -9 | 6. -6 | 21 | 6. | -8 | -6. -6 | 21 | -3. | -7 | 8. -6 |
| 55 | 24 | 6. | -10 | -2. -6 | 23 | -7. | -9 | -2. -6 | 23 | 1. | -7 | 2. -6 |
| 60 | 26 | 2. | -10 | 7. -7 | 25 | -1. | -9 | 6. -7 | 25 | -2. | -8 | 5. -7 |
| 65 | 28 | 3. | -11 | -2. -7 | 28 | 4. | -10 | 2. -7 | 27 | 2. | -9 | -3. -7 |
| 70 | 31 | -2. | -12 | 7. -8 | 30 | 6. | -12 | -6. -8 | 29 | -5. | -10 | -7. -8 |
| 75 | 33 | -6. | -13 | -2. -8 | 32 | -3. | -11 | -3. -8 | 31 | -2. | -10 | -2. -8 |
| 80 | 35 | -8. | -14 | 8. -9 | 34 | 7. | -12 | 7. -9 | 33 | 1. | -10 | 9. -9 |
| 85 | 37 | -2. | -14 | -3. -9 | 36 | -1. | -12 | 3. -9 | 36 | 2. | -11 | 2. -9 |
| 90 | 39 | -4. | -15 | 8. -10 | 38 | 2. | -13 | -7. -10 | 38 | 5. | -13 | 6. -10 |
| 95 | 42 | -6. | -16 | -3. -10 | 40 | -1. | -14 | -3. -10 | 40 | -1. | -12 | -3. -10 |
| 100 | 44 | -7. | -16 | 9. -11 | 42 | -9. | -15 | 8. -11 | 42 | 3. | -13 | -8. -11 |

Table 4

Unnormazlied Arithmetic as an Accuracy Monitor

Computation of sin $(\frac{\pi}{4} + \frac{\pi}{2}n)$

Column A is the number of bits in the sine using unnormalized arithmetic.

Column B is the number of accurate bits in the sine using unnormalized arithmetic.

Column C is the number of accurate bits in the sine using normalized arithmetic.

| n | A | B | C |
|---|---|---|---|
| 0 | 54 | 53 | 53 |
| 1 | 52 | 50 | 50 |
| 2 | 50 | 48 | 54 |
| 3 | 48 | 47 | 47 |
| 4 | 46 | 44 | 44 |
| 5 | 44 | 42 | 46 |
| 6 | 42 | 40 | 39 |
| 7 | 40 | 39 | 40 |
| 8 | 37 | 35 | 34 |
| 9 | 35 | 35 | 33 |
| 10 | 33 | 32 | 34 |
| 11 | 31 | 30 | 29 |
| 12 | 29 | 28 | 29 |
| 13 | 26 | 26 | 25 |
| 14 | 24 | 21 | 21 |
| 15 | 22 | 19 | 20 |
| 16 | 20 | 20 | 17 |
| 17 | 18 | 18 | 17 |
| 18 | 15 | 13 | 17 |
| 19 | 13 | 10 | 11 |
| 20 | 11 | 10 | 10 |
| 21 | 9 | 9 | 9 |
| 22 | 6 | 4 | 4 |
| 23 | 4 | 3 | 4 |
| 24 | 0 | 0 | 4 |

## References


[1] Pucci, Carlo. Sui problemi di Cauchy non "ben posti." Rend. dell'Accad. Naz. dei Lincei (Classe de Sc. fis. mat. e nat.), serie VIII, v. 18 (1955), pp. 473-477.

[2] Pucci, Carlo. Discussone del problema di Cauchy per le equazioni di tipo ellittico. Annali di mat. pura ed applicata, serie IV, v. 46 (1958), pp. 131-154.

[3] Picone, Mauro. Sul calcolo delle fuzioni olomorfe di una variabile complessa. Studies in Mathematics and Mechanics presented to Richard von Mises (1954), pp. 118-126.

[4] John, F. Numerical solution of the equation of heat conduction for preceding times. Annali di mat. pura ed applicata, serie IV, v. 40 (1955), pp. 129-142.

[5] Lin, C. C. and Rubinov, S. I. On the Flow Behind Curved Shocks. J. Math. and Phys., v. 27, n. 2 (1948), pp. 105-129.

[6] Richtmyer, R. D. Detached-Shock Calculations by Power Series, I. AEC Research and Development Report, NYO-7973 (1957), Inst. of Math. Sci., NYU. (This report may be obtained by writing to the AEC Computing and Applied Mathematics Center Library, 4 Washington Pl., New York 3, New York.)

[7] Van Dyke, M. D. A Model of Supersonic Flow Past Blunt Axisymmetric Bodies, with Application to Chester's Solution. J. Fluid Mech., v. 3, pt. 5 (1958), pp. 515-522.

[8] Van Dyke, M. D. The Supersonic Blunt-Body Problem -- Review and Extension. J. Aero/Space Sciences, v. 25, n. 8 (1958), pp. 485-496.

[9] Titchmarsh, E.C. The Theory of Functions. Oxford University Press, Second Edition (1939), pp. 32.

[10] Bochner, S. and Martin, W. T. Several Complex Variables. Princeton Univ. Press (1948), pp. 30-32.

[11] Garabedian, P. R. and Lieberstein, H. M. On the Numerical Calculation of Detached Bow Shock Waves in Hypersonic Flow. J. of Aeronautical Sci., v. 25, n. 2 (1958), pp. 109-118.

MAR 2 1960

Date Due

| | | | |
|---|---|---|---|
| FEB 1 9 1970 | | | |
| | | | |
| | | | |
|

C. 1

NYU
NYO-
2881 Lewis.
Two methods using power

C. 1

C. 1

NYU
NYO-
2881 Lewis.

Two methods using power
TITLE series for solving analytic
initial value problems.

DATE DUE | BORROWER'S NAME | ROOM NUMBER

**N. Y. U. Institute of Mathematical Sciences**
25 Waverly Place
New York 3, N. Y.